ORC 69-27
SEPTEMBER 1969

AVERAGE COST SEMI-MARKOV DECISION PROCESSES

by

SHELDON M. ROSS

OPERATIONS RESEARCH CENTER

COLLEGE OF ENGINEERING
UNIVERSITY OF CALIFORNIA · BERKELEY

20

AVERAGE COST SEMI-MARKOV DECISION PROCESSES

by

Sheldon M. Ross
Department of Industrial Engineering
and Operations Research
University of California, Berkeley

SEPTEMBER 1969 ORC 69-27

This research has been supported by the U. S. Army Research Office-Durham under Contract DA-31-124-ARO-D-331 with the University of California. Reproduction in whole or in part is permitted for any purpose of the United States Government.

ABSTRACT

The Semi-Markov Decision model is considered under the criterion of long-run average cost. A new criterion, which for any policy considers the limit of the expected cost incurred during the first n transitions divided by the expected length of the first n transitions, is considered. Conditions guaranteeing that an optimal stationary (non-randomized) policy exist are then presented. It is also shown that the above criterion is equivalent to the usual one under certain conditions.

AVERAGE COST SEMI-MARKOV DECISION PROCESSES

by

Sheldon M. Ross

1. INTRODUCTION

A process is observed at time 0 and classified into some state $x \in \mathcal{X}$. After classification, an action $a \in A$ must be chosen. Both the state space $\mathcal{X}$ and the action space A are assumed to be Borel subsets of complete, separable metric spaces.

If the state is x and action a is chosen, then

(i) the next state of the process is chosen according to a known regular conditional probability measure $P(\cdot \mid x,a)$ on the Borel sets of $\mathcal{X}$, and

(ii) conditional on the event that the next state is y, the time until the transition from x to y occurs is a random variable with known distribution $F(\cdot \mid x,a,y)$. After the transition occurs, an action is again chosen and (i) and (ii) are repeated. This is assumed to go on indefinitely.

We further suppose that a cost structure is imposed on the model in the following manner: If action a is chosen when in state x and the process makes a transition t units later, then the cost incurred by time $s (s \leq t)$ after the action was taken is given by a known real-valued Baire function $C(s \mid x,a)$.[†]

[†] If one allows the cost to also depend upon the next state visited, then $C(s \mid x,a)$ should be interpreted as an expected cost.

In order to ensure that transitions do not take place too quickly, we shall need to assume the following:

<u>Condition 1</u>:

There exists $\delta > 0$, $\varepsilon > 0$, such that

$$\int_{y \in X} F(\delta \mid x,a,y)dP(y \mid x,a) < 1 - \varepsilon \qquad \text{for all } x , a .$$

In other words, Condition 1 asserts that for every state x and action a there is a positive probability of at least ε that the transition time will be greater than δ .

A policy π is any measurable rule for choosing actions. The problem is to choose a policy which minimizes the expected average cost per time. When the time between transitions is identically 1 , then the process is called a Markov decision process and has been extensively studied (see, for instance, [2], [5] and [6]). When this restriction is lifted, we have a semi-Markov decision process and results have only previously been given for the case where A and S are finite (see [3] and [4]).

2. EQUALITY OF CRITERIA

Let X_n and a_n be respectively the nth state of the process and the nth action chosen, $n = 1,2, \ldots$ Also, let τ_n be the time between the $(n-1)$st and the nth transition, $n \geq 1$.

Furthermore, let $Z(t)$ denote the total cost incurred by t, and let Z_n be the cost incurred during the nth transition interval;† and define for any policy π

$$\phi_\pi^1(x) = \overline{\lim_{t\to\infty}}\, E_\pi\left[\frac{Z(t)}{t} \mid X_1 = x\right]$$

and

$$\phi_\pi^2(x) = \overline{\lim_{n\to\infty}}\, \frac{E_\pi\left[\sum_{i=1}^{n} Z_i \mid X_1 = x\right]}{E_\pi\left[\sum_{i=1}^{n} \tau_i \mid X_1 = x\right]} .$$

Thus ϕ^1 and ϕ^2 both represent, in some sense, the average expected cost. Though ϕ^1 is clearly more appealing, it will be criterion ϕ^2 that we shall deal with. Fortunately, it turns out that under certain conditions both criterions are identical.

<u>Definition</u>:

A policy is said to be stationary if the action it chooses only depends on the present state of the system.

The reader should note at this point that if a stationary policy is employed then the process $\{X(t), t \geq 0\}$ is a semi-Markov process, where $X(t)$ represents the state of the process at time t.

†Of course, $Z(t)$ and Z_n are determined by X_i, a_i, τ_i, $i \geq 1$.

For any initial state x, let

$$T = \inf \{t > 0 : X(t) = x \, , \, X(t^-) \neq x\} \, ,$$

and

$$N = \min \{n > 0 : X_{n+1} = x\} \, .^{\dagger}$$

Hence, T is the time of the first return to state x and N is the number of transitions that it takes.

<u>Lemma 1</u>:

If Condition 1 holds, and if $E_\pi[T \mid X_1 = x] < \infty$, then $E_\pi[N \mid X_1 = x] < \infty$ and $T = \sum_{n=1}^{N} \tau_n$.

<u>Proof</u>:

By the definition of T and N it follows that $T \geq \sum_{n=1}^{N} \tau_n$, with equality holding if $N < \infty$. Now, if we let

$$\bar{\tau}_n = \begin{cases} 0 \text{ if } \tau_n \leq \delta \\ \delta \text{ with probability } \dfrac{\varepsilon}{\int_y (1 - F(\delta \mid x,y,a))dP(y \mid x,a)} & \text{if } \tau_n > \delta \, , \; X_n = x \, , \, a_n = a \\ 0 \text{ with probability } 1 - \dfrac{\varepsilon}{\int_y (1 - F(\delta \mid x,y,a))dP(y \mid x,a)} & \text{if } \tau_n > \delta \, , \; X_n = x \, , \, a_n = a \end{cases}$$

then it follows from Condition 1 that $\bar{\tau}_n$, $n = 1,2, \ldots$ are independent and identically distributed with

†If the set in brackets is empty then take N to be ∞, and similarly for T.

$$P\{\bar{\tau}_n = \delta\} = \varepsilon = 1 - P\{\bar{\tau}_n = 0\} .$$

Now, from Wald's equation it follows that if $EN = \infty$ then $E\sum_1^N \bar{\tau}_n = \infty$, and hence that $ET \geq E\sum_1^N \tau_n \geq E\sum_1^N \bar{\tau}_n = \infty$ (since $\bar{\tau}_n \leq \tau_n$) .

Q.E.D.

Theorem 1:

Assume Condition 1. If π is a stationary policy, and if $E_\pi[T \mid X_1 = x] < \infty$, then

$$\phi_\pi^1(x) = \phi_\pi^2(x) = \frac{E_f[Z(T) \mid X_1 = x]}{E_f[T \mid X_1 = x]} .$$

Proof:

Suppose throughout the proof that $X_1 = x$. Now, under a stationary policy $\{X(t), t \geq 0\}$ is a regenerative process with regeneration (or cycle) point T . Hence, by a well known result

$$\phi_\pi^1(i) = E_\pi[\text{cost incurred during a cycle}]/E_\pi[\text{length of cycle}]$$

$$= E_\pi[Z_T]/E_\pi T .$$

Also, it is easy to see that $\{X_n , n = 1,2, \ldots\}$ is a discrete time regenerative process with regeneration time N . Hence, by regarding $Z_1 + \ldots + Z_N$ as the "cost" incurred during the first cycle of this process, it follows by the same well known result that

(1) $$E_\pi \sum_{n=1}^m Z_n/m \to E_\pi \sum_{n=1}^N Z_n/E_\pi N \qquad \text{as } m \to \infty ,$$

where we have used Lemma 1 to assert that $E_\pi N < \infty$. However, we may also regard $\tau_1 + \dots + \tau_N$ as the "cost" incurred during the first cycle and hence, by the same reasoning,

(2) $$E_\pi \sum_{n=1}^{m} \tau_n/m \to E_\pi \sum_{n=1}^{N} \tau_n / E_\pi N \qquad \text{as} \quad m \to \infty .$$

By combining (1) and (2) we obtain

$$\phi_\pi^2(x) = \frac{E_\pi \sum_{n=1}^{N} Z_n}{E_\pi \sum_{n=1}^{N} \tau_n} .$$

However, since $N < \infty$ (Lemma 1) it is easy to see that $\sum_{n=1}^{N} Z_n = Z(T)$ and $\sum_{n=1}^{N} \tau_n = T$, and the result follows.

Q.E.D.

<u>Remarks</u>:

It also follows from the above proof that, with probability 1,

$$\lim_{t\to\infty} \frac{Z(t)}{t} = \lim_{m\to\infty} \frac{\sum_{n=1}^{m} Z_n}{\sum_{n=1}^{m} \tau_n} = \frac{E_\pi[Z(T)]}{E_\pi T} .$$

Also, suppose that the initial state is y , $y \neq x$. When is it true that $\phi_\pi^1(y) = \phi_\pi^2(y) = \phi_\pi^1(x)$? One answer is that if, with probability 1, the process will eventually enter state x , then $\{X(t), t \geq 0\}$ is a delayed (or general) regenerative process, and the proof goes through in an identical manner.

Let

$$\bar{\tau}(x,a) = \int_{y \in \mathcal{X}} \int_0^{\infty} t dF(t \mid x,a,y) dP(y \mid x,a)$$

and

$$\bar{C}(x,a) = \int_{y \in \mathcal{X}} \int_0^{\infty} C(t \mid x,a) dF(t \mid x,a,y) dP(y \mid x,a) .$$

We shall suppose that both $\bar{C}(x,a)$ and $\bar{\tau}(x,a)$ exist and are finite for all x , a .

We also note that the expected cost incurred during a transition interval and the expected length of a transition interval only depend on the parameters of the process through $\bar{\tau}(x,a)$, $\bar{C}(x,a)$ and $P(\cdot \mid x,a)$; and, as a result, ϕ^2 only depends on the parameters of the process through these three functions. Thus, we may choose the cost and transition time distributions in as convenient a manner as possible; and hence for the remainder of this paper, let us suppose that

$$F(t \mid x,a,y) = \begin{cases} 1 & t \geq \bar{\tau}(x,a) \\ 0 & t < \bar{\tau}(x,a) \end{cases}$$

and

$$C(t \mid x,a) = \begin{cases} 0 & t < \bar{\tau}(x,a) \\ \bar{C}(x,a) & t \geq \bar{\tau}(x,a) . \end{cases}$$

That is, we suppose that the time until transition is (with probability 1) $\bar{\tau}(x,a)$ and that a cost of $\bar{C}(x,a)$ is incurred at the time of transition.

3. AVERAGE COST RESULTS

Theorem 2:

Assuming Condition 1, if there exists a bounded Baire function $f(x)$, $x \in \mathcal{X}$, and a constant g, such that

$$(3) \qquad f(x) = \min_a \left\{ \bar{C}(x,a) + \int_{\mathcal{X}} f(y)dP(y \mid x,a) - g\bar{\tau}(x,a) \right\} \qquad x \in \mathcal{X},$$

then, for any policy π^* which, when in state x, selects an action minimizing the right side of (3), we have

$$\phi^2_{\pi^*}(x) = g = \min_\pi \phi^2_\pi(x) \qquad \text{for all } x \in \mathcal{X}.$$

Proof:

Let $S_i = (X_1,a_1, \ldots, X_i,a_i)$, $i = 1,2, \ldots$ For any policy π

$$E_\pi\left[\sum_{i=2}^{n} [f(X_i) - E_\pi(f(X_i) \mid S_{i-1})]\right] = 0 .$$

But,

$$\begin{aligned}
E_\pi[f(X_i) \mid S_{i-1}] &= \int_{\mathcal{X}} f(y)dP(y \mid X_{i-1},a_{i-1}) \\
&= \bar{C}(X_{i-1},a_{i-1}) + \int_{\mathcal{X}} f(y)dP(y \mid X_{i-1},a_{i-1}) - g\bar{\tau}(X_{i-1},a_{i-1}) \\
&\qquad - \bar{C}(X_{i-1},a_{i-1}) + g\bar{\tau}(X_{i-1},a_{i-1}) \\
&\geq \min_a \left\{ \bar{C}(X_{i-1},a) + \int_{\mathcal{X}} f(y)dP(y \mid X_{i-1},a) - g\bar{\tau}(X_{i-1},a) \right\} \\
&\qquad - \bar{C}(X_{i-1},a_{i-1}) + g\bar{\tau}(X_{i-1},a_{i-1}) \\
&= f(X_{i-1}) - \bar{C}(X_{i-1},a_{i-1}) + g\bar{\tau}(X_{i-1},a_{i-1}) ,
\end{aligned}$$

with equality for π^* , since π^* takes the minimizing actions. Hence,

$$0 \leq E_\pi \sum_{i=2}^{n} [f(X_i) - f(X_{i-1}) + \bar{C}(X_{i-1},a_{i-1}) - g\bar{\tau}(X_{i-1},a_{i-1})]$$

or

$$g \leq \frac{E_\pi \sum_{i=2}^{n} \bar{C}(X_{i-1},a_{i-1})}{E_\pi \sum_{i=2}^{n} \bar{\tau}(X_{i-1},a_{i-1})} + \frac{E_\pi[f(X_n) - f(X_1)]}{E_\pi \sum_{i=2}^{n} \bar{\tau}(X_{i-1},a_{i-1})} ,$$

with equality for π^* . By letting $n \to \infty$ and using the boundedness of f and the fact that Condition 1 implies that $E_\pi \sum_{1}^{n} \bar{\tau}(X_{i-1},a_{i-1}) \geq n\,\varepsilon\,\delta \to \infty$, we obtain

$$g \leq \overline{\lim_{n\to\infty}} \frac{E_\pi \sum_{i=2}^{n} \bar{C}(X_{i-1},a_{i-1})}{E_\pi \sum_{i=2}^{n} \bar{\tau}(X_{i-1},a_{i-1})} = \phi_\pi^2(X_1)$$

with equality for π^* and for all possible values of X_1 .

Remarks:

The above proof is an adaptation of one given in [6] for Markov decision processes. We have tacitly assumed that a rule minimizing the right side of (3) may be chosen in a measurable manner. Clearly a sufficient (but by no means necessary) condition is that the action space A be countable.

In order to determine sufficient conditions for the existence of a bounded function $f(x)$ and a constant g satisfying (3), we introduce a discount factor α , $0 < \alpha < \infty$, and continuously discount costs. That is, we suppose that

a cost of C incurred at time t is equivalent to a cost $Ce^{-\alpha t}$ incurred at time 0.

Let $V_{\pi,\alpha}(x)$ denote the total expected discounted cost when π is employed, and the initial state is x; and let $V_\alpha(x) = \inf_\pi V_{\pi,\alpha}(x)$. Then, it may be shown by standard arguments (see [1]) that

$$V_\alpha(x) = \min_a \left\{ e^{-\alpha\bar{\tau}(x,a)} \left[\bar{C}(x,a) + \int_0^\infty V_\alpha(y)dP(y \mid x,a) \right] \right\}. \tag{4}$$

Now, fix some state--call it 0--and define

$$f_\alpha(x) = V_\alpha(x) - V_\alpha(0).$$

From (4), we obtain

$$V_\alpha(0) + f_\alpha(x) = \min_a \left\{ e^{-\alpha\bar{\tau}(x,a)} \left[\bar{C}(x,a) + \int_0^\infty f_\alpha(y)dP(y \mid x,a) + V_\alpha(0) \right] \right\}. \tag{5}$$

We shall need the following condition:

<u>Condition 2</u>:

There exists an $M < \infty$, such that

$$\bar{C}(x,a) \leq M\bar{\tau}(x,a) \qquad \text{for all } x, a.$$

Theorem 3:

Under Conditions 1 and 2, if the action space A is finite, and if $\{f_\alpha(x)\ ,\ 0 < \alpha < c\}$ is a uniformly bounded equicontinuous family of functions for some $0 < c < \infty$, then

(i) there exists a bounded continuous function $f(x)$ and a constant g satisfying (3);

(ii) for some sequence $\alpha_n \to 0$, $f(x) = \lim_{n\to\infty} f_{\alpha_n}(x)$;

(iii) $\lim_{\alpha\to 0} \alpha V_\alpha(x) = g$ for all $x \in \mathcal{X}$.

Proof:

From (5), we obtain that

$$(6)\qquad f_\alpha(x) = \min_a \left\{ e^{-\alpha\bar{\tau}(x,a)} \left[\bar{C}(x,a) + \int_0^\infty f_\alpha(y)\,dP(y \mid x,a) \right] - V_\alpha(0)(\alpha\bar{\tau}(x,a) + o(\alpha)) \right\} .$$

Now, by the Arzela-Ascoli theorem there exists a sequence $\alpha_n \to 0$ and a continuous function f such that $\lim_{n\to\infty} f_{\alpha_n}(x) = f(x)$ for all x. Also, it follows from Conditions 1 and 2 that $\alpha V_\alpha(0)$ is bounded, and hence we can require that $\lim_{n\to\infty} \alpha_n V_{\alpha_n}(0) \equiv g$ exists. The results (i) and (ii) then follow by letting $\alpha_n \to 0$ in (6) and using Lebesgue's dominated convergence theorem.

The proof of (iii) is identical with the one given in [6].

4. AN EXAMPLE

Suppose that batches of letters arrive at a post office at a Poisson rate λ . Suppose further that each batch consists of j letters with probability P_j , $j \geq 1$, independently of each other. At any time, a truck may be dispatched to deliver the letters. Assume that the cost of dispatching the truck is K , and also that the cost rate when there are j letters present is C_j , an increasing, positive, bounded sequence, $j \geq 1$. The problem is to choose a policy minimizing the long-run average cost.

The above may be regarded as two action semi-Markov decision process with states $1,2,3, \ldots$; where state i means that there are i letters presently in the post office. Action 1 is "dispatch a truck" and action 2 is "don't dispatch a truck." (Note that since a truck would never be dispatched if there were no letters in the post office, we need not have a state 0 .)

The parameters of the process are:

$$P(j/i,1) = P_j \quad , \; P(i + j/i,2) = P_j$$

$$\bar{\tau}(i,1) = 1/\lambda \quad , \; \bar{\tau}(i,2) = 1/\lambda$$

$$\bar{C}(i,1) = K + \frac{C(0)}{\lambda} \; , \; \bar{C}(i,2) = \frac{C(i)}{\lambda} \; .$$

Now, if we let

$$e^{\alpha/\lambda} V_\alpha(i,1) = \min\left\{K + \frac{C(0)}{\lambda} \; ; \; \frac{C(i)}{\lambda}\right\} \; ,$$

and for $n > 1$

$$e^{\alpha/\lambda} V_\alpha(i,n) = \min\left\{K + \frac{C(0)}{\lambda} + \sum_{j=1}^{\infty} P_j V_\alpha(j,n-1) \; ; \; \frac{C(i)}{\lambda} + \sum_{j=1}^{\infty} P_j V_\alpha(i+j,n-1)\right\} \; ,$$

then it follows by induction that $V_\alpha(i,n)$ is increasing in i for each n . Also, since costs are bounded and the discount factor $e^{-\alpha/\lambda} < 1$, it follows that

$V_\alpha(i) = \lim_n V_\alpha(i,n)$, and hence $V_\alpha(i)$ is increasing. Also, $V_\alpha(i)$ satisfies

$$(7) \quad e^{\alpha/\lambda} V_\alpha(i) = \min \left\{ K + \frac{C(0)}{\lambda} + \sum_{j=1}^{\infty} P_j V_\alpha(j) \ ; \ \frac{C(i)}{\lambda} + \sum_{j=1}^{\infty} P_j V_\alpha(i+j) \right\} .$$

We will now show that $V_\alpha(i) - V_\alpha(1)$ is uniformly bounded and hence Theorem 3 is applicable. To do this, we consider two cases:

<u>Case i</u>:

$$e^{\alpha/\lambda} V_\alpha(1) = K + \frac{C(0)}{\lambda} + \sum_{j=1}^{\infty} P_j V_\alpha(j) .$$

In this case, we have by (7) that $V_\alpha(i) \leq V_\alpha(1)$ and hence, by monotonicity,

$$V_\alpha(i) = V_\alpha(1) \qquad \text{for all } i .$$

<u>Case ii</u>:

$$e^{\alpha/\lambda} V_\alpha(1) = \frac{C(1)}{\lambda} + \sum_{j=1}^{\infty} P_j V_\alpha(1+j) .$$

In this case, we have by (7) that

$$\begin{aligned} e^{\alpha/\lambda} V_\alpha(1) \leq e^{\alpha/\lambda} V_\alpha(i) &\leq K + \frac{C(0)}{\lambda} + \sum_{j=1}^{\infty} P_j V_\alpha(j) \\ &\leq K + \frac{C(0)}{\lambda} + \sum_{j=1}^{\infty} P_j V_\alpha(j+1) \\ &= K + \frac{C(0)}{\lambda} - \frac{C(1)}{\lambda} + e^{\alpha/\lambda} V_\alpha(1) . \end{aligned}$$

Thus, in either case $V_\alpha(i) - V_\alpha(1)$ is uniformly bounded and hence by Theorem 3 there exists an increasing function $f(i)$ and a constant g such that

$$f(i) = \min\left\{K + \frac{C(0)}{\lambda} + \sum_{j=1}^{\infty} P_j h(j) - \frac{g}{\lambda} \, ; \, \frac{C(i)}{\lambda} + \sum_{j=1}^{\infty} P_j h(j + i) - \frac{g}{\lambda}\right\},$$

and the policy which chooses the minimizing actions is optimal.

Now, if we let

$$i^* = \min\left\{i : \frac{C(i)}{\lambda} + \sum_{j=1}^{\infty} P_j h(j + i) > K + \frac{C(0)}{\lambda} + \sum_{j=1}^{\infty} P_j h(j)\right\},$$

then it follows from the monotonicity of $C(i)$ and $h(i)$ that the optimal policy is to dispatch a truck whenever the number of letters in the post office is at least i^* ; and hence, the structure of the optimal policy is determined.

REFERENCES

[1] Blackwell, D., "Discounted Dynamic Programming," Ann. Math. Statist., Vol. 36, pp. 226-235, (1965).

[2] Derman, C., "Denumerable State Markovian Decision Processes - Average Cost Criterion," Ann. Math. Statist., Vol. 37, pp. 1545-1554, (1966).

[3] Howard, R., "Semi-Markovian Decision Processes," Int. Stat. Inst. Bull., (1963).

[4] Jewell, W. S., "Markov Renewal Programming I and II," Opns. Res., Vol. 2, No. 6, pp. 938-971, (1963).

[5] Ross, S. M., "Non-Discounted Denumerable Markovian Decision Models," Ann. Math. Statist., Vol. 39, pp. 412-423, (1968).

[6] Ross, S. M., "Arbitrary State Markovian Decision Processes," Ann. Math. Statist., Vol. 39, pp. 2118-2122, (1968).

Unclassified
Security Classification

DOCUMENT CONTROL DATA - R & D

(Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified)

1. ORIGINATING ACTIVITY (Corporate author) University of California, Berkeley	2a. REPORT SECURITY CLASSIFICATION Unclassified 2b. GROUP

3. REPORT TITLE

AVERAGE COST SEMI-MARKOV DECISION PROCESSES

4. DESCRIPTIVE NOTES (Type of report and inclusive dates)

Research Report

5. AUTHOR(S) (First name, middle initial, last name)

Sheldon M. Ross

6. REPORT DATE	7a. TOTAL NO. OF PAGES	7b. NO. OF REFS
September 1969	15	6

8a. CONTRACT OR GRANT NO. DA-31-124-ARO-D-331 b. PROJECT NO. 20014501B14C c. d.	9a. ORIGINATOR'S REPORT NUMBER(S) ORC 69-27 9b. OTHER REPORT NO(S) (Any other numbers that may be assigned this report)

10. DISTRIBUTION STATEMENT

This document has been approved for public release and sale; its distribution is unlimited.

11. SUPPLEMENTARY NOTES NONE	12. SPONSORING MILITARY ACTIVITY U.S. Army Research Office-Durham Box CM, Duke Station Durham, North Carolina 27706

13. ABSTRACT

SEE ABSTRACT.

DD FORM 1 NOV 65 1473 (PAGE 1)
S/N 0101-807-6811

Unclassified
Security Classification

A-31408

Unclassified

Security Classification

14 KEY WORDS	LINK A		LINK B		LINK C	
	ROLE	WT	ROLE	WT	ROLE	WT
Semi-Markov Decision Process						
Average Cost Criterion						
Equality of Criteria						
Semi-Markov Process						
Discounted Cost Criterion						

DD FORM 1 NOV 65 1473 (BACK)

S/N 0101-807-6821

Unclassified

Security Classification

A-31409